D942–D948 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 31 October 2012*
doi:10.1093/nar/gks1012

CellLineNavigator: a workbench for cancer cell line analysis

Markus Krupp[1,2], Timo Itzel[1], Thorsten Maass[1], Andreas Hildebrandt[2], Peter R. Galle[1] and Andreas Teufel[1,*]

[1]Department of Medicine I of the Johannes Gutenberg University Mainz, Langenbeck Strasse 1, 55131 Mainz, Germany and [2]Institute of Computer Science of the Johannes Gutenberg University Mainz, Staudingerweg 9, 55128 Mainz, Germany

Received August 15, 2012; Revised September 24, 2012; Accepted October 2, 2012

ABSTRACT

The CellLineNavigator database, freely available at http://www.medicalgenomics.org/celllinenavigator, is a web-based workbench for large scale comparisons of a large collection of diverse cell lines. It aims to support experimental design in the fields of genomics, systems biology and translational biomedical research. Currently, this compendium holds genome wide expression profiles of 317 different cancer cell lines, categorized into 57 different pathological states and 28 individual tissues. To enlarge the scope of CellLineNavigator, the database was furthermore closely linked to commonly used bioinformatics databases and knowledge repositories. To ensure easy data access and search ability, a simple data and an intuitive querying interface were implemented. It allows the user to explore and filter gene expression, focusing on pathological or physiological conditions. For a more complex search, the advanced query interface may be used to query for (i) differentially expressed genes; (ii) pathological or physiological conditions; or (iii) gene names or functional attributes, such as Kyoto Encyclopaedia of Genes and Genomes pathway maps. These queries may also be combined. Finally, CellLineNavigator allows additional advanced analysis of differentially regulated genes by a direct link to the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics Resources.

INTRODUCTION

In vitro cancer cell culture experiments provide the opportunity of analysing and modelling the complex mechanisms of tumour biology through facile experimental manipulations, global as well as detailed mechanistic studies. They are, therefore, of significant aid in molecular biomedical research. A crucial role of cancer cell lines for medical, scientific and pharmaceutical institutions was elucidated by systematic analysis on lung cancer cell lines (1,2). They revealed not only the amazingly complex role of the cancer genome but also identified and characterized driver mutations in those cell lines. Further studies on cancer cell lines lead to the characterization of tumor protein 53 (TP53) and the understanding of multiple genetic mutations, mutant allele-specific imbalances and copy number losses in cancer (3–5). Moreover, the ability to translate these findings to clinical applications had led to rational therapeutic drug selection (6). For example, activating mutations in the epidermal growth factor receptor (EGFR) kinase domain have major clinical implications in lung cancer, and it was shown in cell line experiments that tumours with this mutation are sensitive to tyrosine kinase inhibitors (7). However, repeatedly a varying response to treatment or targeted manipulation of gene expression was observed in diverse cancer cell lines. This was attributed to a diverse genetic background and, subsequently, a diverse gene expression. Thus, information on these diverse gene expression profiles in cancer cell lines may be crucial to experimental designs of modelling cancer *in vitro* and testing for novel therapeutic approaches.

We have, therefore, generated CellLineNavigator, a workbench for the biomedical community, which allows querying the transcriptom of a great variety of cancer cell lines to screen for the most suitable cell line for upcoming experiments. To enlarge the scope of this database, the data were linked to common functional and genetic databases, enabling querying for a more systematic view on cell line expression profiles.

In summary, we have generated a comprehensive database containing expression profiles of 317 cancer cell lines representing 57 different pathological states and 28 individual tissues. This database will aid the design of *in vitro* experiments in cancer research, as it will allow

*To whom correspondence should be addressed. Tel: +49 6131 17 2380; Fax: +49 6131 17 5595; Email: teufel@uni-mainz.de

taking the genetic background of these cell lines into consideration. The CellLineNavigator database is publicly available at http://www.medicalgenomics.org/celllinenavigator/.

MATERIALS AND METHODS

Data source, data processing

Genome-wide expression data of multiple cell lines, freely available at ArrayExpress [database ID: E-MTAB-37 (8)], were publicly provided by Greshock *et al*. (Laboratory of Cancer Metabolism Drug Discovery, GlaxoSmithKline, Collegeville, PA, USA). The cell lines were handled as previously described (9). Briefly, the transcript abundance of 317 cancer cell lines was analysed using the Affymetrix Human Genome-U133 Plus2 GeneChip technology. This chip covers the complete human genome for analysis of >45 000 transcripts and >19 000 genes. All data were available in technical triplicates. Corresponding information on tissue site and disease state was supported for each cell line (Figure 1).

The differential expression was analysed using the R-Project (10)/bioconductor (11) suite with the following additional libraries: 'affy' (12), 'hgu133plus2.db' (13) and 'frma' (14,15). After quality control, two microarray experiments (cell line SNU398—Replicate 1 and cell line SNU423—Replicate 2) were neglected for further analysis because of insufficient RNA level detection. All data were normalized using the 'expresso' function of the 'affy' package and following settings: background adjustment method: 'mas', normalization method: 'quantiles', PerfectMatch (PM) adjustment method: 'mas' and the method used for the computation of expression values: 'medianpolish'. Next, we calculated the expression median for each probe set for all cell lines. These values were subsequently used as control to calculate log2 transformed expression ratios (*M*-values), after the median expression was calculated for each cancer cell line. *M*-values representing the expression levels of tissue sites and disease states were calculated accordingly. Gene expression barcodes were generated using the 'frma' (frozen robust multiarray analysis) (default options) and 'barcode' (output: *Z*-score) function implemented in the 'frma' package. A frma *Z*-score of >5 suggested that a gene is expressed in a particular tissue. The frma *Z*-score was generated to allow comparison of the expression profiles with data already present at medicalgenomics.org (16,17) and other microarray data sets processed with the frma method. Official gene symbols and National Center for Biotechnology Information (NCBI) Entrez GeneIDs were assigned to the data using the 'hgu133plus2.db' package.

To enable an integrative comparison and querying between gene expression and biological function information, all data were linked to commonly used and established bioinformatics databases and knowledge repositories, such as NCBI Entrez database (18), HUGO Gene Nomenclature Committe (HGNC) (19), Human Protein Reference Database (HPRD) (20), Online Mendelian Inheritance in Man (OMIM) (21), BioGPS (22), Nextbio (23) and Gent (24). Moreover, the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (25) was connected to identify gene signalling and molecular pathway associations. Data on cellular component, biological process and molecular function were collected from the Gene Ontology database (26). Finally, the CellLineNavigator was cross-linked to our RNA-Seq expression profiling database on normal tissues, RNA-Seq Atlas (16), and our liver-specific Library of Molecular Associations [LoMA (17)].

Data organization and Webinterface

The backbone of CellLineNavigator is a Linux-Postgre SQL-Apache-PHP stack implemented in a content management system (Drupal: http://drupal.org/). The database organization is founded on a menu, allowing to directly accessing the following sections: news, data, search, download and help section.

Information on current statistics and recent changes were posted in the news section to keep the users up-to-date, whereas the download section provides the possibility to download the complete CellLineNavigator database in tab separated text file format.

CellLineNavigator may easily be accessed through a simple (data section) or advanced querying interface (search section).

The data section offers the possibility to explore tissues (default) or disease states (Figure 2). Filtering options for expression levels within individual tissues or disease states are supported. The default filter is set to list all genes with a different expression level of at least 2-fold in comparison with the respective control. Six additional levels of expression filtering are supported (from 1.5- to 5-fold). However, the user may also set the filter criteria to none (no filter) or no regulation (list all genes whose *M*-values are in the range of −1 to +1).

To allow users a high degree of flexibility to access CellLineNavigator, we implemented an advanced search section, offering the user 'Fulltext search' or 'Explore profile' options (Figure 3). The 'Fulltext search' may be used to query for individual genes provided by the user to query for expression levels within specific cell lines, tissues or disease states (or any combination of all). Using the 'Explore profile' query option, the user may query for specific expression levels classified in the fields of (i) genes (ii) KEGG pathway maps (iii) gene ontologies (iv) cell lines (v) tissues or (vi) disease states. Again, a combination of all query types is possible. Moreover, the user may also define cut-off criteria to filter for specific expression levels. The resulting gene list is shown in an interface providing the same features mentioned in the data section with one exception, the filter criteria is adjusted to the preceding query. These features may again be used for further filtering the resulting gene list.

To allow users a more customizable way of displaying the expression results, an extra option for setting the regulation view is supported (default: 2-fold).

Moreover, a powerful resource within our database extending the full impact of the individual gene–cell line relations is provided in the details section (Figure 4). This

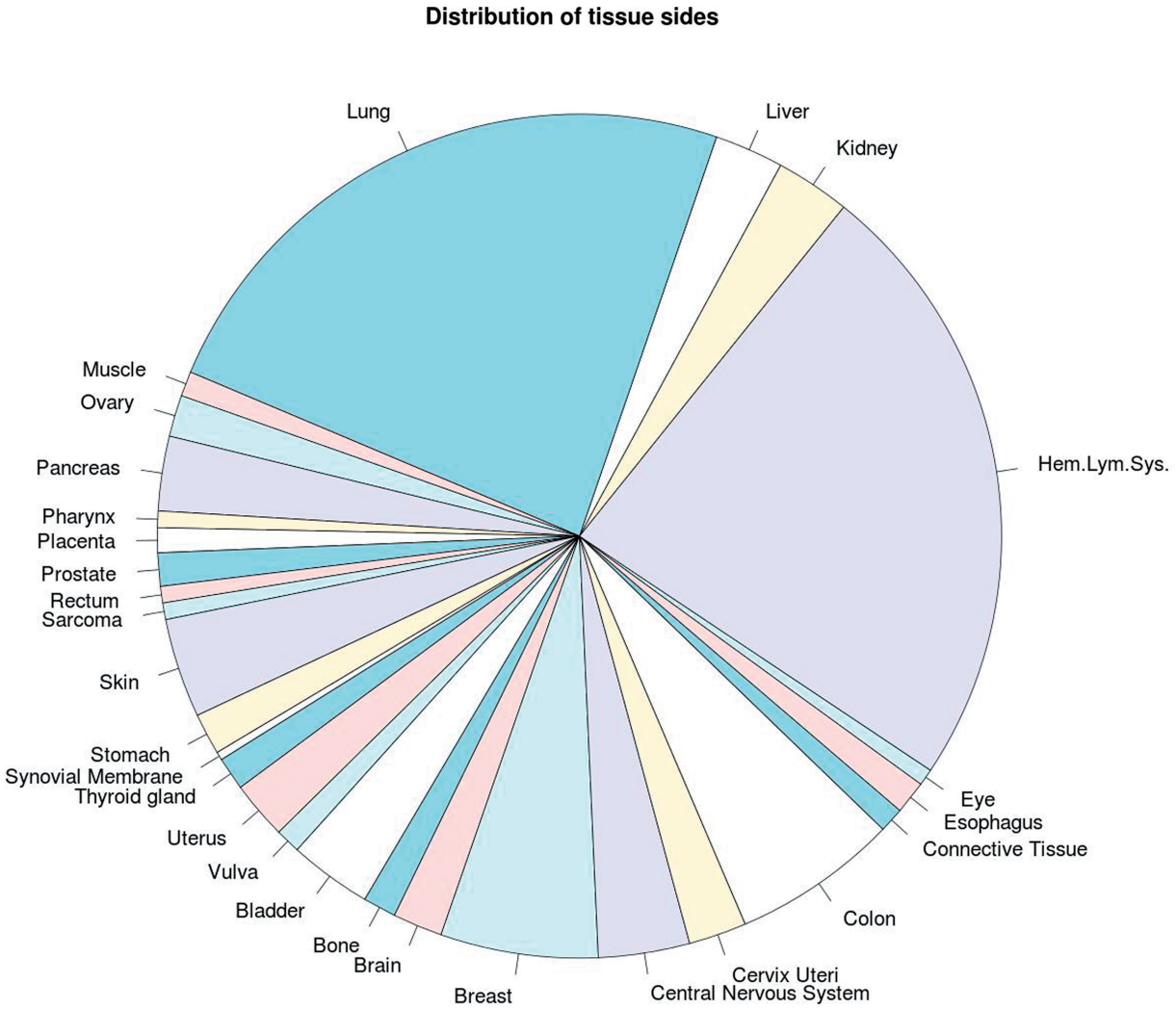

Figure 1. Distribution of tissues within CellLineNavigator.

section may be accessed by either clicking on the detail link or the specific expression icon in the results tables. It provides additional information on gene symbol, description, aliases, chromosomal location, Entrez ID, Ensembl ID, Gene Ontology, KEGG pathway and expression profiles. Although the expression profiles were individualized to the previous user query, for example, did the user click on the expression icon of tissue side 'bladder', the details view will show a barchar with an overview of the expression within all tissues and, more importantly, with a barchar representing the specific expression values of the cell lines corresponding to the tissue of interest. For further comparison with already available data at medicalgenomics.org, such as RNA-Seq Atlas, the details view may be switched from *M*-value to frma *Z*-score representation.

Finally, a major strength of the database is its direct connection to the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics Resources (27). Gene lists generated in CellLine Navigator may automatically be transferred to the DAVID analysis tools.

DISCUSSION

The current scope of biomedical studies on tumorigenesis requires a tremendous amount of human tumour material. Stringent restrictions on the international exchange of biological reagents and increasing requirements from institutes, ethics committees and government are limiting the availability of those human tumour materials. Thus, *in vitro* cell culture is highly useful for modelling the complex mechanisms of cancer development to identify molecular mechanisms related to tumour development and potential therapeutic targets.

Cell lines are capable of infinite replication and, therefore, offer an unlimited source of biomedical material that can be distributed to laboratories worldwide and thus, allow direct comparison of research results if originating from identical material. As a matter of fact, these cell lines are widely used in biomedical research. However, the detailed knowledge about their genetic profiles is still limited and has not been summarized in a large comparative database.

Diverse biological behaviour of cancer cell lines may result from diverse underlying genetic profiles and

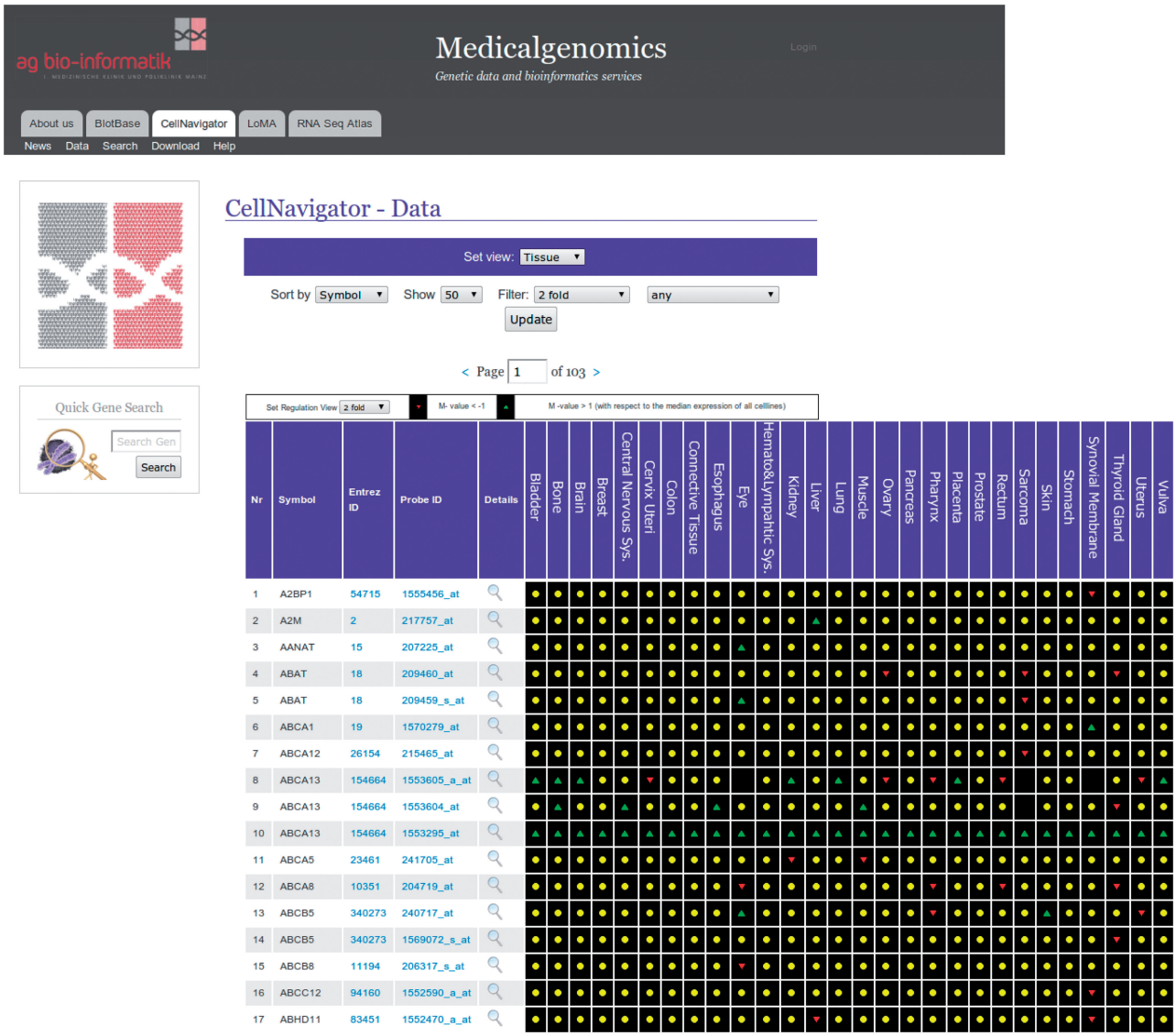

Figure 2. The data section offers the possibility to explore tissues or disease states. In default, the user will get a list of all genes that have shown a different expression level of at least 2-fold in comparison with the respective control within the specific tissues. The set view option allows the user to switch between tissues and disease state views. Additional six filter options for expression levels are supported (default 2-fold). Further, the user can set the expression filter criteria to none (no filter) or no regulation (list all genes whose *M*-values are in the range of −1 to +1). To allow users a more customizable way in displaying the data, the user may change the cut-off criteria for differentially expressed genes (default: 2-fold).

expression signatures, which may differ significantly among immortalized tumour cell lines. The awareness of these differences makes it useful/necessary to take the diverse gene expression signatures into account, especially while planning targeted strategies to influence the biological behaviour. Large scale microarray experiments to unravel the genetic profile of these cell lines are available through public databases, such as ArrayExpress (8) of Gene Expression Omnibus (28).

However, the analysis of these data is hardly feasible for biologists or physicians without substantial bioinformatics skills or at least knowledge on microarray analysis. Even with profound experience in microarray technology, analysis of such data is complex and time consuming task. So far to our knowledge, the Gene Expression Atlas (8) is the only database that provides access to these cell line expression profiles. However, the main focus of this database is not on cancer cell lines, and thus, it just contains the expression profile of ~90 cell lines from various species. Moreover, not only the classification into specific phenotypes but also data collected from multiple laboratories are incomplete and, therefore, exhibit multiple experimental conditions, making a comparison between the multiple expression profiles extremely difficult.

The database, CellLineNavigator, presented here contains gene expression profiles of >300 human cancer cell lines. These expression profiles were generated in the same laboratory under nearly the same experimental conditions and thus, guarantee a highest degree on comparability. Further, depending on phenotypic information, these cell lines were classified into corresponding

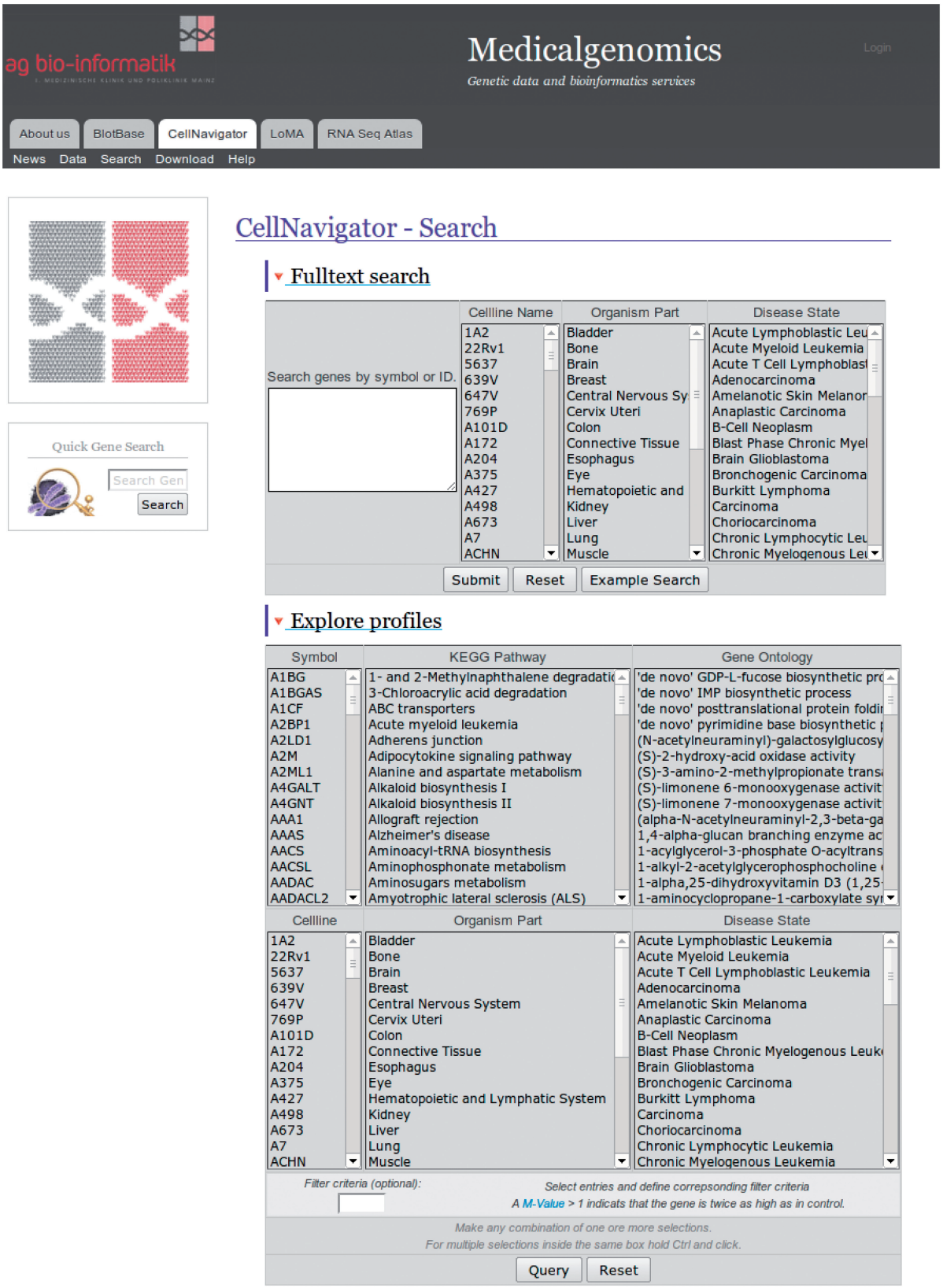

Figure 3. The search section allows users to choose between the 'Fulltext search' or 'Explore profile' option. In the 'Fulltext search', the user can provide a gene list that can be queried for expression levels within specific cell lines, tissue sides or disease states (or any combination of all). The 'Explore Profile' allows the user to query for specific expression levels classified in the fields (i) gene, (ii) KEGG pathway maps, (iii) Gene Ontology, (iv) cell line, (v) tissue side or (vi) disease state. A combination of all query types is possible. Additionally, the user may define a cut-off criteria to filter for specific expression levels.

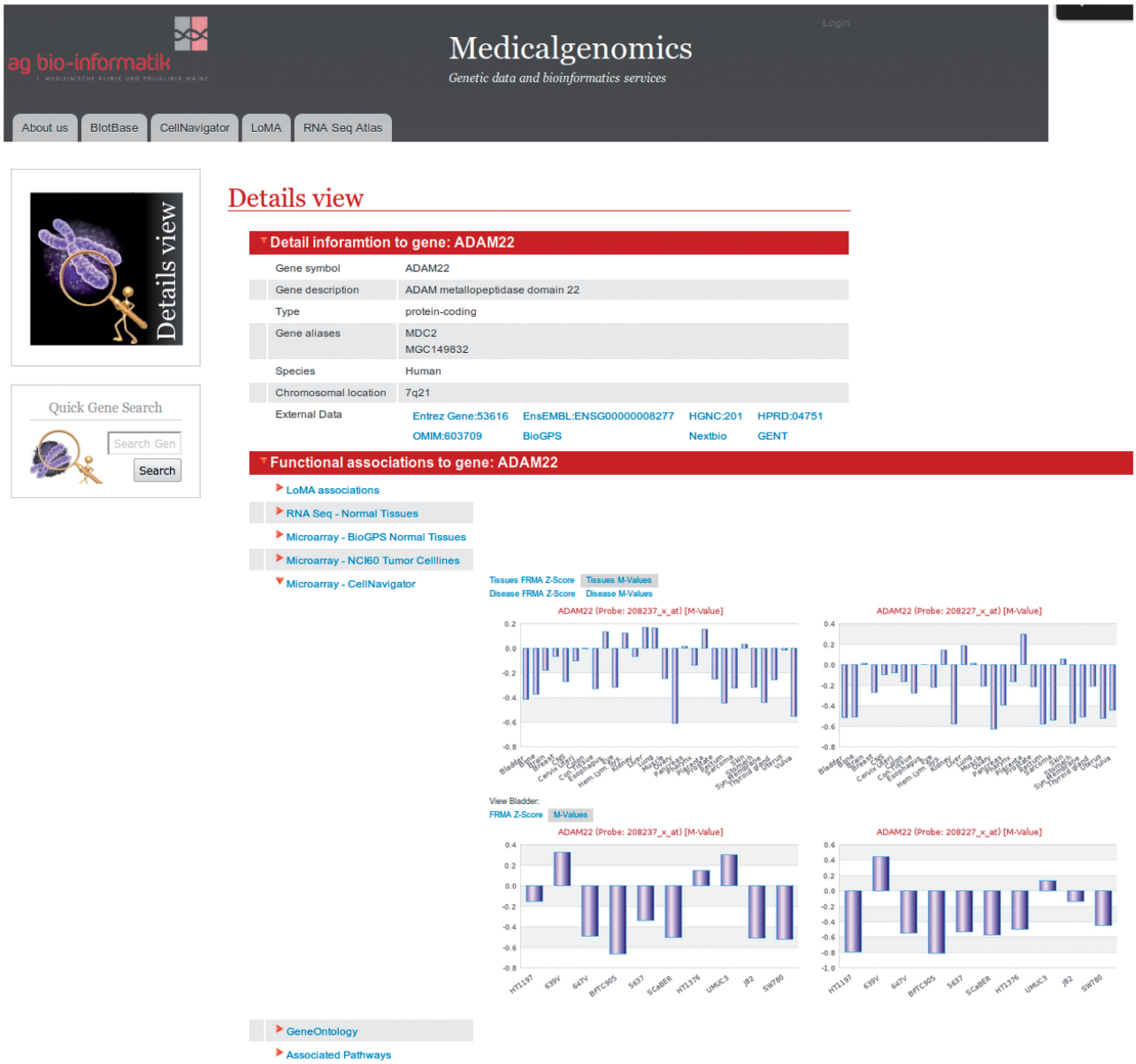

Figure 4. The details section offers a powerful option to access the full impact of the individual gene–cell line relation. The expression profiles are individualized to the previous user query, for example, if the user is interested in the expression of the tissue 'bladder', the details view display a barchar showing an overview of the expression within all tissues and, more importantly, with a barchar representing the specific expression values of the cell line(s) corresponding to the to the tissue 'bladder'. Further, additional information on gene symbol, description, aliases, chromosomal location, Entrez ID, Ensembl ID, Gene Ontology and KEGG pathway are supported.

tissues of origin and disease states. The main focus of CellLineNavigator is not simply on summarizing these data but rather on an easy and user friendly availability as well as the linkage to advanced bioinformatics analyses tools. To guarantee easy data access and connectivity, we implemented a mostly self-explaining Web application as a user friendly front end to the data base. This Web application allows users to query for (i) differentially expressed genes; (ii) pathological (e.g. melanoma) or physiological (e.g. lung) conditions or (iii) gene names or functional attributes, such as KEGG pathway maps. A combination of all query types is possible.

Comparative analysis of differential gene expression between cell lines or diseases of interest will initially often result in (large) lists of genes being differentially regulated. To further characterize the differences between the respective samples and thus a major advance in the usability of this database, these collections of genes need to be further characterized with respect to functional or structural similarities. We, therefore, chose to link and provide an automated data transfer of gene lists of interest to DAVID, a large bioinformatics suite providing functional and structural analyses, such as pathway enrichment, gene ontology enrichment or analysis of functional domains. This automated linkage to DAVID brings our database resource and analysis tool to the next level of not only comparing genetic changes but also functionally and structurally

characterizing the differences by means of advanced bioinformatics. In summary, CellLineNavigator is the first database providing comprehensive summary, display and analysis options for gene expression data of the most commonly used cancer cell lines. It provides access to large microarray data sets without advanced bioinformatics skills. Thus, CellLineNavigator may be of significant aid for *in vitro* modelling of cancer mechanisms and testing of novel therapeutic approaches.

FUNDING

Boehringer Ingeheim Funds and Roche (to AT). Funding for open access charge: Department of Medicine I of the Johannes Gutenberg University Mainz.

Conflict of interest statement. None declared.

REFERENCES

1. Pleasance,E.D., Cheetham,R.K., Stephens,P.J., McBride,D.J., Humphray,S.J., Greenman,C.D., Varela,I., Lin,M.-L., Ordóñez,G.R., Bignell,G.R. *et al.* (2009) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.
2. Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature*, **458**, 719–724.
3. Gandhi,J., Zhang,J., Xie,Y., Soh,J., Shigematsu,H., Zhang,W., Yamamoto,H., Peyton,M., Girard,L., Lockwood,W.W. *et al.* (2009) Alterations in genes of the EGFR signaling pathway and their relationship to EGFR tyrosine kinase inhibitor sensitivity in lung cancer cell lines. *PLoS One*, **4**, e4576.
4. Singh,A., Greninger,P., Rhodes,D., Koopman,L., Violette,S., Bardeesy,N. and Settleman,J. (2009) A gene expression signature associated with 'K-Ras addiction' reveals regulators of EMT and tumor cell survival. *Cancer Cell*, **15**, 489–500.
5. Soh,J., Okumura,N., Lockwood,W.W., Yamamoto,H., Shigematsu,H., Zhang,W., Chari,R., Shames,D.S., Tang,X., MacAulay,C. *et al.* (2009) Oncogene mutations, copy number gains and mutant allele specific imbalance (MASI) frequently occur together in tumor cells. *PLoS One*, **4**, e7464.
6. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
7. Paez,J.G., Jänne,P.A., Lee,J.C., Tracy,S., Greulich,H., Gabriel,S., Herman,P., Kaye,F.J., Lindeman,N., Boggon,T.J. *et al.* (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, **304**, 1497–1500.
8. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al.* (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.
9. Greshock,J., Bachman,K.E., Degenhardt,Y.Y., Jing,J., Wen,Y.H., Eastman,S., McNeil,E., Moy,C., Wegrzyn,R., Auger,K. *et al.* (2010) Molecular target class is predictive of in vitro response profile. *Cancer Res.*, **70**, 3677–3686.
10. R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
11. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
12. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
13. Carlson,M., Falcon,S., Pages,H. and Li,N. hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2).
14. McCall,M.N., Bolstad,B.M. and Irizarry,R.A. (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.
15. McCall,M.N., Uppal,K., Jaffee,H.A., Zilliox,M.J. and Irizarry,R.A. (2011) The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res.*, **39**, D1011–D1015.
16. Krupp,M., Marquardt,J.U., Sahin,U., Galle,P.R., Castle,J. and Teufel,A. (2012) RNA-Seq Atlas—a reference database for gene expression profiling in normal tissue by next generation sequencing. *Bioinformatics*, **28**, 1184–1185.
17. Buchkremer,S., Hendel,J., Krupp,M., Weinmann,A., Schlamp,K., Maass,T., Staib,F., Galle,P.R. and Teufel,A. (2010) Library of molecular associations: curating the complex molecular basis of liver diseases. *BMC Genomics*, **11**, 189.
18. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
19. Seal,R.L., Gordon,S.M., Lush,M.J., Wright,M.W. and Bruford,E.A. (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.*, **39**, D514–D519.
20. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
21. Amberger,J., Bocchini,C. and Hamosh,A. (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum. Mutat.*, **32**, 564–567.
22. Wu,C., Orozco,C., Boyer,J., Leglise,M., Goodale,J., Batalov,S., Hodge,C., Haase,J., Janes,J., Huss,J. *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
23. Kupershmidt,I., Su,Q.J., Grewal,A., Sundaresh,S., Halperin,I., Flynn,J., Shekar,M., Wang,H., Park,J., Cui,W. *et al.* Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS One*, **5**, e13066.
24. Shin,G., Kang,T.-W., Yang,S., Baek,S.-J., Jeong,Y.-S. and Kim,S.-Y. (2011) GENT: gene expression database of normal and tumor tissues. *Cancer Inform.*, **10**, 149–157.
25. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
26. Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
27. Jiao,X., Sherman,B.T., Huang,D.W., Stephens,R., Baseler,M.W., Lane,H.C. and Lempicki,R.A. (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**, 1805–1806.
28. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.